REPORT R-349 MAY, 1967

CSL COORDINATED SCIENCE LABORATORY

ON CLUSTERING TECHNIQUES OF CITATION GRAPHS

F.P. PREPARATA
R.T. CHIEN

UNIVERSITY OF ILLINOIS – URBANA, ILLINOIS

This work was supproted in part by the Joint Services Electronics Program (U. S. Army, U. S. Navy, and U. S. Air Force) under Contract No. DA 28 043 AMC 00073(E); and in part by NSF GK-690.

ON CLUSTERING TECHNIQUES OF CITATION GRAPHS

F. P. Preparata and R. T. Chien

ABSTRACT

In this paper we report results in the application of graph theory to the problem of clustering in document retrieval systems using bibliographic coupling devices. The problem is attacked by mapping the citation graph of the document collection onto a unidimensional storage array. The figure of merit of the location assignment is the total distance between connected pairs of documents, or, equivalently, the "stretching" resulting from the mapping. This is the objective function of the problem. An algorithm is then presented for the reduction of the objective function, which provides a currently improving solution. Its computational complexity only grows as $N^{3/2}$, where N is the collection size.

I. Introduction

The problem of organizing a large universe of objects with the purpose to identify sets, in such a fashion that objects within a set are similar to each other but are dissimilar from objects outside the set has received considerable attention over the past years[1,2,3] as a fundamental topic in the theory of classification. As it has been observed in many of the mentioned works, however, the vagueness of terms such as "similar", dissimilar", or, equivalently, the qualitative nature of the relations existing among the objects of the universe have largely prevented the use of a mathematical framework in the modeling of the problem. Yet the notions of similarity and dissimilarity are quite primitive in our semantics and, therefore, organizational criteria inspired by these concepts appear quite natural for large universes of elements.

The above mentioned qualitative nature of the interrelations among objects is also reflected by the adoption of the term "cluster" in lieu of set, thus implying the intuitive identification of some "core" along with some "fuzziness" in the definition of the boundaries of such sets (1).

The "clustering problem" is definitely central in information retrieval, particularly in document retrieval with reference both to document classification and to automatic indexing (see, e.g.[5]). The clustering techniques proposed heretofore[1,2,3] are based on some reasonably defined concept of "cohesion" among members of the document cluster.

---

[1] A closely related concept, in fact, is that of "fuzzy set", proposed by Zadeh[4] with reference to a universe whose elements have various degrees of membership in several sets of the universe.

Given a quantitative value to the pairwise association among documents (for example, based on the number of common keywords) the universe is represented as an undirected graph (undirected because the association between two documents is obviously reciprocal), whose nodes are representative of documents and whose weighted edges are representative of document associations. The reader is referred to[1,2,3] for a detailed discussion of different clustering techniques, all based however on the criterion of assigning a document to the cluster with which it has the highest "global" association. It suffices here to point out that from a computational point of view the proposed methods are characterized by the fact that the effort required grows roughly with the square of the collection size (as one would intuitively expect from methods entirely based on matrix algorithms).

The approach we present in this paper, although closely germane to those mentioned above, draws its immediate motivation from a rather important problem which typically manifests itself in computer-based document retrieval systems. The complexity of an information retrieval task (processing of a query against a file) depends largely on the physical location of the documents in the file. Quite generally, in a computer based system the processing time is a monotone increasing function of the total time necessary to access the item required from the computer storage; each individual access time is in turn a monotone nondecreasing function of the relative distance, in the memory structure, of each pair of items sequentially accessed. From this general remark, it appears quite desirable to locate physically close in the memory structure items that are likely to

be wanted together (for example, in the same cylinder of a disc file or in the same strip of a magnetic strip file).

This aspect of a computer based system becomes dominant when the interrelation among documents is expressed by the relation of citation between a source document and a reference document. In this case, in fact, the search algorithm itself proceeds along paths of a graph, and a means to improve the system's performance is to bring at a small physical distance in storage documents which are close in some intuitively acceptable sense in the collection graph (namely, a "citation" graph). Therefore, the existence of a citation link between a pair of documents is taken as a sign of similarity, or, equivalently, of likelihood of them being wanted together.

In the sequel we discuss a method which is aimed at the identification of sets of documents which are "close" in the citation graph. This is done by mapping the graph onto a unidimensional array and by successively rearranging the locations assigned to document. The criterion governing the location assignment is the reduction of the "stretching" of graph links as produced by the mapping. This is equivalent to the reduction of the total stretching (objective function) and will, on the average, bring to close-by locations in the array documents which are close in the graph. The presented algorithm is effective in the sense that only reductions of the total stretching are produced; in a slightly modified version, it is efficient since its complexity from a computational standpoint grows only as $N^{3/2}$, where N is the size of the collection.

It is interesting to notice that the objective function of the problem is monotonically non-increasing as the proposed algorithm proceeds: hence, depending upon considerations of policy or of diminishing return, processing may be stopped at any point, the resulting configuration being certainly not worse than the initial one.

II. Definitions and Problem Statement

A collection of N documents $\{d_1, d_2, \ldots, d_n\}$ is described as an undirected citation graph, the nodes of which are the documents. An edge ℓ_{hk} between d_h and d_h exists if and only if either 1) "d_h cites d_k" or 2)" d_h is cited by d_k". Hence $\mathcal{D}$ is completely described by its N X N connection matrix $B = \| b_{hk} \|$ where $b_{hk} = b_{kh} = 1$ if and only if either 1) or 2) hold (by this we also imply that all citation edges have the same weight).

With $\mathcal{U}$ we denote a unidimensional array of N cells $\{1,2,\ldots,N\}$, which can be pictured as the set of points with positive integral abscissa on the line segment $[1, N]$.

An assignment A of $\mathcal{D}$ is a mapping of $\mathcal{D}$ onto $\mathcal{U}$ such that if $(j_1, j_2, \ldots, j_n)$ is a permutation of the integers $(1,2,\ldots,N)$, document d_{j_i} is assigned to cell i. $(d_{j_i} \rightarrow i)$.

Given a generic assignment A, assume $b_{hk} = 1$ and that $d_h \rightarrow i_h$ $d_k \rightarrow i_k$. The quantity $s_{hk} = | i_h - i_k |$ is termed the relative stretching of edge ℓ_{hk} under the assignment A. Hence, for each assignment the quantity

$$S = \frac{1}{2} \sum_{h,k=1}^{N} b_{hk} s_{hk}$$

termed the total relative stretching, is perfectly defined and computable.

At this point it is convenient to introduce some functions which are defined on the set of cells of $\mathcal{U}$. To avoid confusion, the value that a function f takes at a cell j will be indicated with f^j (superscripted).

Let $d_{j_i} \rightarrow i$ under an assignment A. Further let n_{j_i} be the degree

of d_{j_i} in $\mathcal{A}$, i.e. the number of documents directly connected to d_{j_i}. Of these, assume that under A, r^i have been assigned cells whose markings are greater than i and ℓ^i have been assigned cells whos markings are smaller than i; therefore

$$n_{j_i} = r^i + \ell^i$$

In other words, r^i, ℓ^i are the numbers of "stretched" edges emanating from i and going respectively to the right and to the left of i, if cells 1,2,...,N are arranged in natural order from left to right.

For each cell i we introduce the <u>incremental</u> function s^i

$$(1) \quad s^i = r^i - \ell^i$$

and the <u>cumulative</u> function

$$(2) \quad f^i = \sum_{j=1}^{i} s^j$$

We notice on passing that f^i gives the number of links that are intercepted by an ideal section between i and i + 1. In fact

$$f^i = \sum_{j=1}^{i} s^j = \sum_{j=1}^{i} r^j - \sum_{j=1}^{i} \ell^j$$

which shows that f^i equals the number of links going to the right from cells 1,2,...,i minus the subset of these which terminate on cells of the set 2,3,...,i, which confirms our assertion. Further

$$(3) \quad S = \sum_{j=1}^{N-1} f^j$$

In fact, $s_{j_h, j_k} = | h - k |$ can be thought of as giving a unit contribution to f^h, f^{h+1},...,f^{k-1} (in the case that $h < k$). From this observation, and the remark that $f^N = 0$ (no links are present on the right of cell N) relation (3) follows immediately.

Example: Assume that the undirected graph of Figure 1 is given.

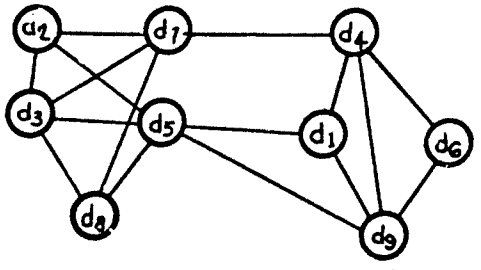

Fig. 1 - A Citation Graph

If now d_j is mapped into cell j of a unidimensional array U we have the following assignment (Figure 2).

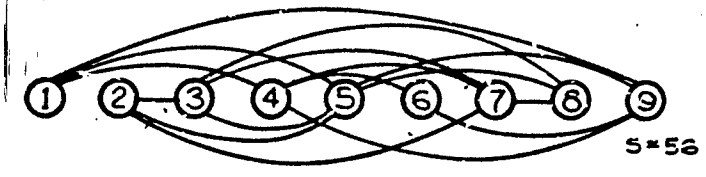

Fig. 2 - Initial Assignment $A \rightarrow U$

For this initial assignment we have a total stretching S = 56.

Let us now consider a generic node d_{j_i}, such that under A $d_{j_i} \rightarrow i$.

Let d_{j_i} be connected to $d_{j_{h_1}}$, $d_{j_{h_2}}, \ldots, d_{j_{h_s}}$ under A, as usual, $d_{j_{h_m}} \rightarrow h_m$. We define the potential function of d_{j_i} as

$$\varphi_i^j = \sum_{r=1}^{s} |j - h_r|$$

In other words φ_i^j gives the sum of the relative stretching of all the links connected to d_{j_i} if d_{j_i} is placed in cell j without affecting the assignment of any other document. We remark that for a given assignment A:

a) for any cell j such that $h_r \leq j < h_{r+1}$ the increment of φ_i^j is given by

$$\varphi_i^{j+1} - \varphi_i^j = r - (s-r)$$

i.e. the difference between the number of links connected to

$d_{j_{h_{r+1}}}, \ldots, d_{j_{h_s}}$ and the number of links connected to $d_{j_{h_1}}, \ldots, d_{j_{h_r}}$.

In fact the displacement of d_{j_i} one position to the right causes the stretchings of the links connected to the former set to increase by one unit, while the ones pertaining to the latter set are decreased by one unit. Hence in the interval $h_r \leq j < h_{r+1}$, φ_i^m is a linear function whose increment is 2r-s (constant in the interval).

b) At each cell $h_1, h_2, \ldots, h_s$ the increment of φ_i^j for increasing m undergoes a positive discontinuity of 2, since at each such cell one link passes from the right set to the left set.

Hence φ_i^j decreases if $2r-s < 0$, increases if $2r-s > 0$. Remarks a) and b) can be combined in the following proposition.

Proposition - If d_{j_i} is connected to $d_{j_{h_1}}, d_{j_{n_2}}, \ldots, d_{j_{h_s}}$ the function φ_i^j is a convex piecewise linear function which attains its minimum at:

$h_{s/2} \leq j \leq h_{s/2+1}$ if s is even

$h_{\frac{s+1}{2}} = j$ if s is odd

As a second remark, following directly from the definitions, we have that

$$S = \frac{1}{2} \sum_{i=1}^{N} \varphi_i^i$$

Finally we consider the problem of generating a new assignment A' from a given assignment A. The basic operation we shall use to this end is

the right cyclic permutation: if $(d_{j_r}, d_{j_{r+1}}, \ldots, d_{j_s})$ are assigned to $(r, r+1, \ldots, s)$ respectively, after performing the cyclic permutation $(s \mid r)$ they will be assigned to $(r+1, r+2, \ldots, s, r)$ respectively. Obviously, any assignment can be obtained from any other assignment through a finite number of right cyclic permutations: in fact any assignment is a permutation, each permutation is equivalent to a finite number of transpositions, each transposition is equivalent to a finite number of right cyclic permutations (RCP).

It is now of interest to find an expression for the change of S determined by an RCP. We first notice that an RCP $(s \mid r)$ results from the successive performance of the dislocation $(d_{j_r} \to r+1, d_{j_{r+1}} \to r+2, \ldots, d_{j_{s-1}} \to s)$ and of the insertion $d_{j_s} \to r$. Let us examine separately the effect of these two operations on S.

Consider the dislocation and the following sets of links:

$\mathcal{T}_{rs}$ = set of links from $(1, 2, \ldots, r-1)$ to $(s, s+1, \ldots, N)$

$\mathcal{U}_{rs}$ = set of links from $(1, 2, \ldots, r-1)$ to $(r, r+1, \ldots, s-1)$

$\mathcal{V}_{rs}$ = set of links from $(r, r+1, \ldots, s-1)$ to $(s, s+1, \ldots, N)$

Let t_{rs}, u_{rs}, v_{rs} be the cardinalities of $\mathcal{T}_{rs}, \mathcal{U}_{rs}, \mathcal{V}_{rs}$ respectively. The dislocation does not affect the stretchings of links of $\mathcal{T}_{rs}$, while the stretchings of all links of $\mathcal{U}_{rs}$ is increased by one unit and of all those of $\mathcal{U}_{rs}$ is decreased by one unit. Hence the change Δs_1 of S due to the dislocation alone is

$$\Delta S_1 = u_{rs} - v_{rs}$$

But since

$$f^{r-1} = u_{rs} + t_{rs}$$

we have $$f^{s-1} = v_{rs} + t_{rs}$$

$$\Delta S_1 = f^{r-1} - f^{s-1}$$

Consider next the insertion $d_{j_s} \rightarrow r$. The change ΔS_2 of S due to this operation alone would be exactly $\varphi_s^r - \varphi_s^s$ if the dislocation did not take place. A correction is therefore necessary. Specifically the stretching of each link from s to $(r, r+1, \ldots, s-1)$ appears reduced by 1 in ΔS_1, while it must actually increase by 1 by effect of the RCP. Hence, if there are ν_{sr} such links the total change of S is

$$\Delta S = \Delta S_1 + \Delta S_2 + 2\nu_{sr} =$$

$$= (f^{r-1} + \varphi_s^r) - (f^{s-1} + \varphi_s^s) + 2\nu_{sr}$$

This is summarized by the following theorem

Theorem - The change ΔS of S determined by the RCP $(s|r)$ is

$$\Delta S_{sr} = (f^{r-1} + \varphi_s^r) - (f^{s-1} + \varphi_s^s) + 2\nu_{sr} \tag{4}$$

where ν_{sr} is the number of links from s to $(r, r+1, \ldots, s-1)$.

After performing and RCP $(s|r)$ the values of f^j are modified. This modification, however affects f^j only for $r \leq j < s$. Specifically let $\nu_{rs} = h$ and, let s be linked to $i_1, i_2, \ldots, i_h$ with $r \leq i_1 < i_2 < \ldots < i_h < s$. Denote with f'^j the values of f^j after performing $(s|r)$. Then we have the following relations:

$$f'^{j} = f^{j} \quad \text{for } 1 \le j < r \text{ and } s < j \le N$$

$$f'^{j} = f^{j-1} + (f^{s} - f^{s-1}) + 2h \qquad r \le j \le i_1$$

$$f'^{j} = f^{j-1} + (f^{s} - f^{s-1}) + 2(h-m) \qquad i_m < j \le i_{m+1} \tag{5}$$

$$(m = 1,2,\ldots,h-1)$$

$$f'^{j} = f^{j-1} + (f^{s} - f^{s-1})$$

$$i_h < j < s$$

We have now all the necessary tools for the development of an algorithm aimed at the reduction of the function S. In fact, if the function f^j is known we can rapidly ascertain from (4) whether a proposed RCP $(s|r)$ will result in a net decrease of S: $\Delta s < 0$ will be assumed as the decision rule for its execution. Secondly, the function f^j can be updated in a relatively simple manner with the aid of relations (5). The algorithm is presented in the next section.

III. An Algorithm for the Reduction of S

Given an assignment A of $\mathfrak{D}$, i.e. a mapping $d_{j_i} \rightarrow i$ where d_{j_i} is a document of the collection and i a cell of U, we construct the following tables of N entries:

a) T_1 - Its i-th entry contains the cell of U in which d_i is currently stored.

b) T_2 - Its j-th entry contains the identification of the document currently stored in cell j of U. (i.e. T_2 is the inverse of T_1).

c) T_3 - Its j-th entry contains the current value of f^j.

d) T_4 - Its h-th entry contains the list of all documents linked in $\mathfrak{D}$ to d_h.

With the aid of these four tables we can now give the following algorithm for the reduction of the total relative stretching S.

Algorithm 1

1. - Set j = 2
2. - Access the j-th entry of T_2: let this be d_h
3. - Access the h-th entry of T_4 and obtain all documents linked to d_h.
4. - Obtain from T_1 the cells in which the documents obtained in step 3 are stored.
5. - With the aid of T_3, compute $\psi^m = f^{m-1} + \varphi_j^m + 2\nu_{jm}$ for m = i, j-1, j-2, ... Find the minimum of ψ^m. Let this be ψ^r.

6. - Form $\Delta = \psi^j - \psi^r$. If $\Delta \leq 0$ go to step 7. Else perform $(j \mid r)$.

7. - Update T_1 and T_2 and T_3.

8. - If $j = N$, stop. Else replace j with j+1 and return to Step 2.

Application of Algorithm 1 certainly satisfies the requirement that the current value of S be monotonically non increasing. In substance, with Algorithm 1 we scan U from left to right one cell at each step and determine whether the document contained in the last scanned cell can be "brought" to the left through an RCP resulting in a net decrease of S. After completion of a left-to-right scanning of U, a right-to-left scanning is performed, to possibly relocate documents for which a $\Delta S > 0$ was obtained during the former scanning: this completes a processing cycle.

Example: Let us consider again the citation graph $\mathcal{G}$ of Fig. 1 and its initial assignment shown in Fig.2.(S = 56). We perform now a left-to-right pass in the application of algorithm 1 to U. The result of this processing is shown in Fig. 3: we also have S = 34.

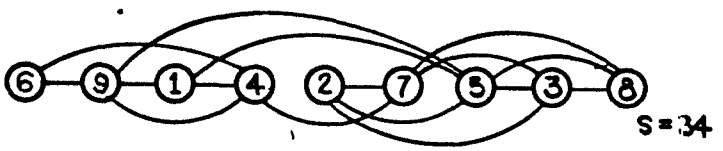

Fig. 3 - After a left-to-right pass

We perform then a right-to-left pass, which yields the assignment shown in Fig. 4 (S = 30).

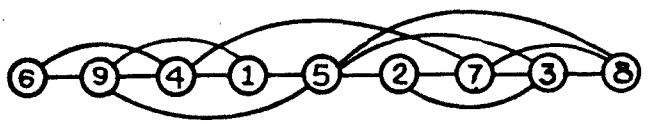

Fig. 4 - After a right-to-left pass

However simple the example may be, the effectiveness of the algorithm is apparent already after a single pass: the two clusters of $\mathcal{G}$ are in fact already identifiable. The performance of a second pass also shows that we are approaching a point of diminishing return in the attempt to reduce S: the application of the algorithm may reasonably stop after obtaining the assignment of Figure 4.

If in a cycle we find that no cyclic permutation can be performed, the processing is terminated: in this case we have reached a local minimum of S. At this point there is no evidence whether the obtained minimum is also the absolute minimum: as a matter of fact it is possible to concoct some clever "interlocking" configurations which correspond to relative mimina. It seems reasonable, however, to introduce at this stage a perturbation of the reached assignment in the form of a single random permutation of the same, and then apply Algorithm 1 to the new assignment. It is likely that this device may lead out of the trap represented by a local minimum.

Leaving this rather important question, some comment is necessary with regard to the computational complexity of Algorithm 1. Steps 2, 3, 4, 6, 8 require a fixed amount of computation per document processed. Step 5, however, requires the calculation of ψ^m for each $m \leq j$ and therefore its complexity is proportional to j. Step 7, in the case that a permutation is performed, requires the updating of T_1, T_2 and T_3, the complexity of which is proportional to $(j-r_j)$ for some $r_j < j$. Let $r_j = \alpha_j\, j$ with $0 < \alpha_j < 1$. Then let $C_i(j)$ be the computational complexity of the i-th step in processing the j-th document. The total complexity C of one scanning is therefore

$$C = \sum_{j=1}^{N} [c_2 + c_3 + c_4 + c_6 + c_8 + c_5 j + c_7 \alpha_j\, j]$$

$$= N(c_2 + c_3 + c_4 + c_6 + c_8 + \frac{c_5}{2}) + \frac{c_5}{2} N^2 + c_7 \sum_{1}^{N} \alpha_j\, j$$

Obviously, for some $0 < \alpha < 1$,

$$\sum_{1}^{N} \alpha_j j = \frac{\alpha}{2} (N^2 + N)$$

hence letting $c_2 + c_3 + c_4 + c_6 + c_8 + \frac{c}{2} 5 + \frac{c_7 \alpha}{2} = a$ and $\frac{c_5}{2} + c_7 \frac{\alpha}{2} = b$

we have

$$C \cong b N^2 + a N$$

which shows a square law rate of growth for C. To avoid this undesirable feature we propose the introduction of some approximations in Algorithm 1. We notice that only steps 5 and 7 contribute to the term in N^2: these we want to modify.

Let us first consider step 7. Assume that a permutation $(u|t)$ has been performed with $u < s$. Without updating the function f^j for $t \leq j < u$ and the location of the documents contained in cells t, t+1,...,u, let us compute the quantity

$$(6) \quad \Delta S^*_{sr} = (\varphi^{*r}_s - \varphi^{*s}_s) + (f^{*r-1} - f^{*s-1}) + 4\nu^*_{sr}$$

(The symbols are asterisked to denote that the functions are relative to the assignment before the permutation $(u|t)$). Then if $t > r$ (See Fig. 5)

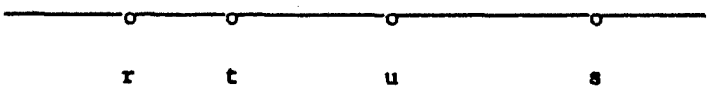

Fig. 5

we have that $f^{s-1} = f^{*s-1}$, $f^{r-1} = f^{*r-1}$ and $\nu_{sr} = \nu^*_{sr}$. Assume then that s is linked to μ cells in the interval (t, t+1,...,u-1), and that $\delta = 1,0$ if s is or is not linked to u, respectively. We have

$$\varphi_s^{*r} = \varphi_s^r - \mu + \delta(u-t+1)$$

$$\varphi_s^{*s} = \varphi_s^s + \mu - \delta(u-t+1)$$

It follows that

$$\Delta S^*_{sr} = (\varphi_s^r - \varphi_s^s) + (f^{r-1} - f^{s-1}) - 2\mu + 2\delta(u-t+1) + 4\nu_{sr}$$

Obviously $\nu_{sr} \geq \mu$ and $\delta(u-t+1) \geq 0$. Hence

$$\Delta S^*_{sr} \geq (\varphi_s^r - \varphi_s^s) + (f^{r-1} - f^{s-1}) + 2\nu_{sr} + 2\delta(u-t+1) \geq \Delta S_{sr}$$

It follows that if $\Delta S^*_{sr} < 0$, then also $\Delta S_{sr} < 0$: our criterion can therefore be applied to ΔS^*_{sr} as given by eq. (6). If $t \leq r$, the same argument can be applied with the only exception that now $f^{r-1} \neq f^{*r-1}$. We notice that due to the permutation $(u|t)$

$$\sum_{j=u}^{t-1} f^j < \sum_{j=u}^{t-1} f^{*j}$$

i.e. the function f^j is decreasing on the average. Hence we feel justified in the conservative approximation $f^{r-1} \simeq f^{*r-1}$, and conclude that

$$\Delta S^*_{sr} \geq \Delta S_{sr}$$

If now ΔS^*_{sr} is computed with reference to the functions which were current before the performance of v permutations, relation (6) is easily generalized to

$$(6) \quad \Delta S^*_{sr} = (\varphi_s^{*r} - \varphi_s^{*s}) + (f^{*r-1} - f^{*s-1}) + 2(v+1)\,\nu^*_{sr}$$

We see therefore that, if a permutation is decided with reference to ΔS^*_{sr}, the updating of T_1, T_2, T_3 can be performed after v executed permutations.

The ensueing updating procedure could be a modification of the one expressed by relations (5): it appears simpler, however to proceed through the reevaluation of s^j for every cell j affected by any of the v permutations. Specifically, let these permutations be $(s_i \mid r_i)$ $(i = 1,2,\ldots,v)$ and let

$$\bar{s} = \max s_i \qquad , \bar{r} = \min r_i \ ;$$

then for $\bar{r} \leq j < \bar{s}$, compute s^j and f^j. Obviously, v must be rather small to avoid gross approximations. A convenient spacing of the updating runs is provided by the following discussion of step 5.

In the search for the minimum of

$$\psi^{*m} = f^{*m-1} + \varphi_j^{*m} + 2(v+1)\ \nu_{jm}$$

after v permutations not followed by updating, assume that the interval (1,j) is subdivided into the following segments:

$$(1,\ a),\ (a+1,\ 2a),\ \ldots\ ,\ (ha+1,\ j)$$

where $h = \lceil \frac{j}{a} \rceil$, i.e. the highest integer smaller than j/a. Let $I_s \equiv [sa+1,\ (s+1)a]$ and $g_s = \min f^{*m-1}$ for $m \in I_s$. Also, let $\varphi_{js} = \max \lceil \varphi_j^{*m} + 2(v+1)\nu_{jm} \rceil$ for $m \in j_s$. It follows that

$$\min \psi^{*m} \leq g_s + \varphi_{js} = \psi_s$$

The function ψ_s is taken as an indication of the values of ψ^{*m} in I_s, the better the approximation the smaller the parameter a. Then if

$$\psi_n = \min \psi_s \qquad\qquad s = 1,2,\ldots,h$$

we perform the calculation of ψ^{*m} for $m \in I_n$.

In summary, the search for ψ^{*m} entails the computation of $\psi_1, \psi_2, \ldots, \psi_h$ and of ψ^{*m} for a different values of m. Hence its complexity is given by

$$c_5\, h + c_5'\, a \simeq c_5\, \frac{j}{a} + c_5'\, a$$

The choice of a as a function of j is crucial with regard to the complexity of step 5 over a complete scanning of the collection. It is very simple to show that the quantity

$$c_5\, \frac{j}{a} + c_5' a$$

for fixed j and variable a attains its lowest value for

$$a \simeq \sqrt{\frac{c_5}{c_5'}}\; j^{1/2}$$

With the insight provided by this relation we subdivide the interval (1,N) into the following set of segments:

$$(1,2),\ (3,6),\ (7,12),\ldots,(p^2 - p+1,\ p^2 + p),\ldots$$

Let $K_p \equiv (p^2 - p+1,\ p^2 + p)$ and p_{max} be the smallest p suchthat $p^2 + p \geq N$ or, approximately, $p_{max} \simeq \sqrt{N}$. For each $i \in K_p$ $a = p$. Hence the computational complexity of step 5 for the totality of $i \in K_p$ is

$$c_5'' \simeq 2(c_5 + c'_5)\ p^2 - (2c_5' + c_5)p$$

which summed over all p's yields approximately

$$\frac{2}{3}\ (c_5 + c_5')\ N^{3/2} - (2c_5' + c_5)\ \frac{N}{2}$$

Analogously, assume that we perform a fixed number w of updating runs per interval K_p. The complexity of an updating run is proportional to p^2, that is

$$c_7' \, p^2$$

which summed over all p's yields

$$\sum_{1}^{j_{max}} c_7' \, p^2 \simeq \frac{c_7'}{3} \, p_{max}^3 \simeq c_7'' \, N^{3/2}$$

In summary, with the artifices introduced to modify the search for the minimum of ΔS and the updating procedures of the pertinent functions, we have a computational procedure whose complexity grows only as $N^{3/2}$.

References

[1] A. F. Parker-Rhodes, "Contribution to the Theory of Clumps", Rep. M.L. 138, Cambridge Language Research Unit, March 1961.

[2] T. T. Tanimoto, "An Elementary Mathematical Theory of Classification and Prediction", International Bus. Mach. Corp. Nov. 1958,10 pgs.

[3] R. E. Bonner, "On Some Clustering Techniques", IBM Jour. Res. Dev., pp. 22-32, January 1964.

[4] L. A. Zadeh, "Fuzzy Sets", Information and Control, Vol. 8, pp. 338-353, 1965.

[5] G. Salton, "Progress in Automatic Information Retrieval",IEEE Spectrum, pp. 90-103, August 1965.

Security Classification

DOCUMENT CONTROL DATA - R & D

Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified

ORIGINATING ACTIVITY (Corporate author)

University of Illinois
Coordinated Science Laboratory
Urbana, Illinois 61801

2a. REPORT SECURITY CLASSIFICATION: Unclassified

2b. GROUP

REPORT TITLE

ON CLUSTERING TECHNIQUES OF CITATION GRAPHS

DESCRIPTIVE NOTES (Type of report and, inclusive dates)

AUTHOR(S) (First name, middle initial, last name)

PREPARATA, F. P. & CHIEN, R. T.

REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
May, 1967	20	5

CONTRACT OR GRANT NO.: DA 28 043 AMC 00073(E) 200014501B31F;
PROJECT NO: also in part NSF GK-690.

9a. ORIGINATOR'S REPORT NUMBER(S): R-349

9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)

DISTRIBUTION STATEMENT

DISTRIBUTION OF THIS REPORT IS UNLIMITED.

SUPPLEMENTARY NOTES

12. SPONSORING MILITARY ACTIVITY: Joint Services Electronics Program thru U.S. Army Electronics Command Fort Monmouth, New Jersey 07703

ABSTRACT

In this paper we report results in the application of graph theory to the problem of clustering in document retrieval systems using bibliographic coupling devices. The problem is attacked by mapping the citation graph of the document collection onto a unidimensional storage array. The figure of merit of the location assignment is the total distance between connected pairs of documents, or, equivalently, the "stretching" resulting from the mapping. This is the objective function of the problem. An algorithm is then presented for the reduction of the objective function, which pfovides a currently improving solution. Its computational complexity only grows as $N^{3/2}$, where N is the collection size.

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Information retrieval						
Clustering						
Document collection						
Citation graph						
Location assignment						
Unidimensional file						
Iterative algorithm						
Computational complexity						